Radboud Universiteit Nijmegen

DONDERS INSTITUTE

Predicting from diversity

How predictions about unknown group members are influenced by relevant and irrelevant group diversity

Marta Blasco Oliver

MSc Thesis Cognitive Neuroscience

On-site supervisors: Lukas Spieß, Prof. Harold Bekkering

Second reader: Prof. Ivan Toni

Student: Marta Blasco Oliver (s1010738)

Date of oral examination: 18th December 2019

Abstract

During social interactions, we are constantly making predictions about other people's behaviour. To inform these predictions, we often make use of what we know about the groups a person belongs to. Previous studies show that the extent to which we use group knowledge to make inferences depends on the diversity within group members. Yet, these studies have not distinguished between group diversity on a property that is relevant to the current goal (i.e. making a prediction about a group member's behaviour), from group diversity on an uninformative, irrelevant property. Therefore, the present study aims to investigate how these two aspects of group diversity influence predictions about unknown group members' behaviour. To do so, we created a novel behavioural paradigm in which group diversity is manipulated both in the goal-relevant dimension (i.e. behaviour) and in a goal-irrelevant dimension (i.e. physical appearance), and participants had to make predictions about the behaviour of unknown group members. As expected, we found that higher goal-relevant diversity leads to more variable and uncertain predictions. Goal-irrelevant diversity also affects predictions in the same direction. Our findings suggest that predictions about unknown group members are not only influenced by group diverstiy in a relevant dimension (i.e. behaviour) but also by diversity in an irrelevant, uninformative dimension of the group (i.e. physical appearance). Finally, possible mechanisms underlying these findings are discussed and further research directions are suggested.

Keywords: Diversity; Group knowledge; Prediction; Category; Variability; Inference

1. Introduction

In daily life situations, we are constantly making predictions about other people's behaviour. To do so, we make use not only of what we know from a specific person but also of what we know about the groups the person belongs to. Previous studies show that the way and the extent to which we use this group knowledge to make predictions about someone depends on how diverse the group they belong is or is perceived to be (e.g. Crawford, Sherman & Hamilton, 2002; Kim & Lee, 2017; Ryan, Judd & Park, 1996; Patalano & Ross, 2007). Yet, these studies have not distinguished between group diversity on a property that is relevant to the current goal (i.e. making a prediction about a group member's behaviour), from group diversity on a property that is not directly relevant to the goal. Therefore, this study aims to investigate how these two aspects of group diversity influence predictions about unknown group members' behaviour.

During everyday life, we are constantly making predictions about surrounding situations. This allows us to respond faster to expected events and overcome ambiguous ones (Clark, 2013). Particularly about other people, we constantly make predictions of how they might behave, what they might prefer or what they believe (Bach & Schenke, 2017). Predictions about people's behaviour are crucial for efficient social interactions. For instance, anticipating a certain behaviour can be useful for coordinating actions with others (Sebanz & Knoblich, 2009), as it is necessary for sports or conversations. Predictions about others' allow us to make decisions and select our behaviour adaptively, as we take into account the consequences of their potential behaviour. There are several sources of information that we use to predict behaviours, preferences or any other property of someone else (Frith & Frith, 2006). One of these sources is individual knowledge: what we know about a particular person, based on for instance previous experiences with them (i.e. Maria is usually late so she will probably be late this time as well). Another common source of information for making predictions is group knowledge (Macrae & Bodenhausen, 2000, Macrae & Bodenhausen, 2001; Frith & Frith, 2006), which is what we know about the groups that a particular person belongs to (e.g. Maria is a teenager, teenagers are usually late, so Maria is likely to be late as well). Group knowledge is often automatically activated and applied when encountering a group member (Allport, 1954; Fiske & Neuberg, 1990; Macrae & Bodenhausen, 2000; Macrae & Bodenhausen, 2001). It also influences the way we perceive, evaluate, judge and even behave towards members of social groups (Axelson et al., 2010; Phelps et al., 2000; Biernat et al., 2009; Green et al., 2007; Ito & Bartholow, 2009; Macrae & Bodenhausen, 2000; Moss-Racusin et al., 2012; Quadflieg et al., 2011; Rachlinski & Guthriet, 2009). Making predictions about someone's behaviour supported by group knowledge can be useful, as it has been shown to be cognitively efficient (Macrae, Milne & Bodenhausen, 1994; see Macrae & Bodenhausen, 2000 for a review). In this fast-paced social world, interactions are constantly happening, thus group knowledge is a simple and fast way to inform inferences about someone's potential behaviour (Fiske & Neuberg, 1990). This anticipations then allow us to behave and take decisions accordingly without the need of having previous experiences with that individual.

Group knowledge is not only about the property that characterizes the group members (e.g. teenagers are usually late) and the magnitude of it (e.g. more or less late), but also about how this property is distributed within a group (e.g. teenagers are all late, but adults are some on time, some a bit late, some very late) (Fried & Holyoak, 1984; Tenenbaum & Griffiths, 2001). Group diversity can be defined as the extent to which members of a group are different from each other, or are perceived as such. In fact, group diversity can influence the way and the strength with which we make inferences about group members based on group knowledge. For instance, it has been shown that the more diverse a social group is perceived to be, the more unlikely it is that group knowledge is applied to a new group member (Patalano & Ross, 2007; Kim & Lee, 2017) and transferred across all group members interchangeably (Crawford et al., 2002). Similarly, higher perception of group diversity leads to being more uncertain about generalizing a group property to an unknown group member (Ryan et al., 1996) and to be more uncertain about members' ambiguous behaviours (Ryan, Bogart & Vender, 2000). Therefore, perceiving a group as diverse seems to lead to a weaker generalisation of group knowledge onto group members.

Notably, when the goal is to make a prediction about a particular property of a group member (e.g. Maria's punctuality), some group properties will be directly relevant and informative to the current goal (e.g. teenagers are usually late), while other properties will be rather irrelevant and uninformative (e.g. teenagers have usually long hair). Based on the property's relevance for the prediction to be made, we distinguish between group diversity in a property that is directly relevant to it (goal-relevant diversity) and group diversity in a property that is not directly relevant (goal-irrelevant diversity). On the one hand, studies that focused on the effect of goal-relevant diversity have shown that, when group variability on a specific trait (i.e. intelligence) is high, we tend to also estimate high variability in that same trait within new group members (Park & Hastie, 1987). Moreover, we tend to consider atypical group members as more plausible (Lambert, 1995), we are less confident about considering a group trait as representative of the group (Lambert et al., 1998) and it is harder to form a global impression of the group (Lambert et al., 1998). Similarly, groups whose members are physically more diverse are judged less accurately in terms of their physical appearance (Dasgupta, Banaji & Abelson, 1999). On the other hand, studies that focused on the effect of goal-irrelevant diversity have shown that, when members of an artificial humanoid group have a more diverse physical appearance (each of them being of a different colour), they are also perceived as psychologically more diverse (Abelson et al., 1998; Dasgupta et al., 1999). They also found that judgments about psychological traits were slower and harder to make for groups of high physical diversity (Dasgupta et al., 1999).

Importantly, earlier research has investigated the effect of goal-relevant and goal-irrelevant group diversity mainly in isolation (Abelson et al., 1998; Dasgupta et al., 1999; Lambert, 1995; Lambert et al., 1998; Park & Hastie, 1987). Social groups in the real world, however, are not only characterized by one property but by several ones, and not all properties are at all times equally relevant or informative for making predictions. Hence previous studies do not allow us to distinguish whether coexisting goal-relevant and goal-irrelevant group diversity influence our predictions in different ways. For example, effects of goal-irrelevant diversity might be overruled by the presence of a goal-relevant group property (as suggested by Goldstone & Yon, 2012); or goal-relevant diversity could influence predictions itself while goal-irrelevant diversity might rather affect the certainty associated to predictions. Moreover, from the studies on goal-irrelevant diversity (e.g., Dasgupta et al., 1999), it is not clear to what extent their findings reflect a genuine influence of group diversity from an uninformative dimension to another one, or whether the effect was induced by task demands. The fact that goal-irrelevant diversity was the only differential information between groups during their experiment, could have made participants become aware of it and subsequently think they had to do something with that information, even if it was actually uninformative for the goal of the task. For these reasons, it is crucial to investigate the effects of goal-relevant and irrelevant group diversity in coexistence.

There are two more aspects that remain unclear from the current literature. First, in most of the previous studies, participants had to make explicit judgments about group members such as reporting a likelihood or answering a question (Abelson et al., 1998; Dasgupta et al., 1999; Kim & Lee, 2017; Lambert, 1995; Park & Hastie, 1987; Patalano & Ross, 2007; Ryan et al., 1996). However, the effects of group diversity on how we make socially meaningful predictions, instead of explicit judgments, has not been addressed. Second, how goal-relevant and/or goal-irrelevant group diversity influence predictions particularly about unknown group members has not been clearly addressed either. That is, studies looking at specific diversity effects only focused on already known members (e.g. Dasgupta et al., 1999) or a whole new group (Park & Hastie, 1987), while the ones that focused on unknown members were based on perceived group diversity (Kim & Lee, 2017; Patalano & Ross, 2007; Ryan et al., 1996). Group members that are already known are associated with both individual and group knowledge. Hence focusing on unknown groups members is crucial to isolate the influence of exclusively group knowledge (and its characteristics, like diversity) on the predictions about group members.

As a consequence of the abovementioned points, the current study aims to investigate the influence of coexisting goal-relevant group diversity and goal-irrelevant group diversity on how we make meaningful predictions about unknown group members' behaviour. In doing so, we will assess (i) how variable the predictions are (ii) the extent to which they diverge from the average group behaviour and (iii) how uncertain those predictions are, as a function of goal-relevant and goal-irrelevant group diversity.

To address these aspects, a novel behavioural paradigm has been designed. Group diversity of several sports teams is experimentally manipulated both in a goal-relevant dimension (in this case, throwing behaviour) and a goal-irrelevant dimension (in this case, physical appearance in terms of skin colour and hair), and predictions about unknown group members' behaviour need to be made. Participants play a computer sports-like game with these teams. Regarding the manipulation of behavioural diversity, some teams throw the ball to a smaller range of locations (low behavioural diversity) while other teams throw the ball to a wider range of locations (high behavioural diversity). Similarly, to manipulate diversity in physical appearance, some teams are composed by members that look very alike (low physical diversity) while others are composed by members that look very different (high physical diversity). Per team, participants first observe where a few team members throw the ball, and then make predictions about the throwing behaviour of members they have not seen throwing. These predictions have a meaningful purpose within the context of the game. Based on previous findings, we hypothesize that participants' predictions about unknown group members' behaviour will be more variable, less based on the average group behaviour and more uncertain for teams where goal-relevant (behavioural) diversity is high, compared to low. Furthermore, we also hypothesize that high goal-irrelevant diversity (i.e. in physical appearance) will additively enhance these effects on predictions about behaviour.

2. Methods

2.1. Participants

Sixty-seven healthy participants took part in our experiment. Two of them were excluded due to a technical error during testing, therefore 65 participants (14 male, M = 23.59 years, SD = 4.49, range 18 to 35 years old) were included in the analysis. Additionally, the same analysis was run without outliers (see criteria in Data Analysis section), for which 62 participants were included (14 male, M = 23.42

years, *SD* = 4.51, range 18 to 35 years old). Participants were recruited via the online research participation system of Radboud University and did not report any neurological disorder. They all gave written informed consent after receiving oral and written information about the experiment. The study was approved by the local ethics committee (ECSW-2018-115) and followed the standards according to the Declaration of Helsinki. As compensation for their participation in the study, participants received either a 7.5 euros voucher or an equivalent of 1 course credit.

2.2. Stimuli

All the instructions and parts of the task ran on a computer (Dell Precision T3610, 4 x 3.7 GHz, 8GB ram) with Windows 7 and presented on an LCD monitor (Benq XL2420Z, 24 inches, 120Hz, 1920x1080 pixels). Responses were given with a Corsair Vengeance K70 keyboard and a Logitech G500 mouse depending on the task instructions of each moment. Due to task manipulations, locations on the screen were defined in polar coordinates instead of cartesian coordinates, with pixels (px) as eccentricity units and degrees (°) as angular units. The computerized task was programmed and run in Python 2.7.12, using Psychopy toolbox (version 1.84.2). All customized stimuli were created and edited with Adobe Photoshop CC 2019 and Illustrator CC 2019.

The task is contextualized in a circular playfield on the screen where team players make throws from the centre to a certain angle and eccentricity (see figure 1A for a schematic depiction of the playfield). The participant's location and movement around the playfield are displayed on real time and depicted as the catcher of the game, a semi-transparent white player (size = 70 x 90 px) with a cross in the centre (size = 30 x 30 px). A baseball-like ball (diameter = 30 px) was the one being thrown by the team players, and whose potential location participants had to predict.

The main stimuli for the study were the team players. These are composed by a face generated with the online Mii Studio from Nintendo (https://studio.mii.nintendo.com/) and a customized team t-shirt, together having a size of 120 x 180 px. A differently customized set of 7 players was used for the practice team. There are a total of 28 teams, thus 28 unique team t-shirts were created. By selecting 2 out of 8 colours in all possible combinations, we created 28 unique bicolour striped patterns), so that each team had a unique t-shirt (see figure 1B). This allows participants to group players of a team together, and distinguish them from the other teams of the game. Each team has 7 players, therefore 196 unique faces were systematically created. For each face, we manipulated 6 features (skin colour, hair colour, hairstyle, eye colour, nose shape and face shape). Each feature had 7 possible phenotypes (as there are 7 members per team). On the one hand, this allowed for the possibility of feature uniqueness within a team. That is, for a certain feature (i.e. eye colour), each team member could have a unique phenotype of it (i.e. one had brown eyes, another green eyes, another blue eyes, etc., without phenotype repetition). On the other hand, this also allowed for the possibility of all members sharing a feature. That is, for a certain feature (i.e. eye colour), all team members could have the same phenotype (i.e. all members green eyes). This system enabled us to change the level of diversity of the teams in terms of physical appearance: members of a high physical diversity team are unique in all features, whereas members of a low physical diversity team share some features (see figure 1B for examples). Notable, by looking at a single player it is not possible to know if it belongs to a low or a high physical diversity teams, as the same 6 features were manipulated in all players, but they were combined in teams following the explained different criteria.

2.3. Procedure and phases of the task

For the task, we developed a sports game called Catch-the-Ball. Participants had to play this game with 28 different teams, each constituted by seven team members. Teams are manipulated in terms of behavioural diversity (low or high) and physical diversity (low or high), as depicted in figure 1. Per team, first, all members are visually introduced (phase 1) and then participants observe four members throwing the ball to certain locations (phase 2). This way, participants are exposed to the physical appearance of all team members and the throwing behaviour of a subset of them. By observing these events, the participant might form an impression (i.e. group knowledge) of the characteristics of the team. Immediately after, they have to make predictions about where six team members might throw the ball, and how certain they are about that prediction (phase 3). Importantly, they have to make predictions about members they have seen throwing the ball in the previous phase (seen players), and about members they have not seen any behaviour from (unseen players). Lastly, participants are asked to judge, for each of those six team members, whether they were seen members (i.e. they saw them throwing during the second phase) or unseen members (phase 4). The structure and the timeline of a complete round are depicted in figure 2. Of note, the framing story of the task, together with the context and the stimuli used, were meant to induce a more immersive and interactive sensation to participants.

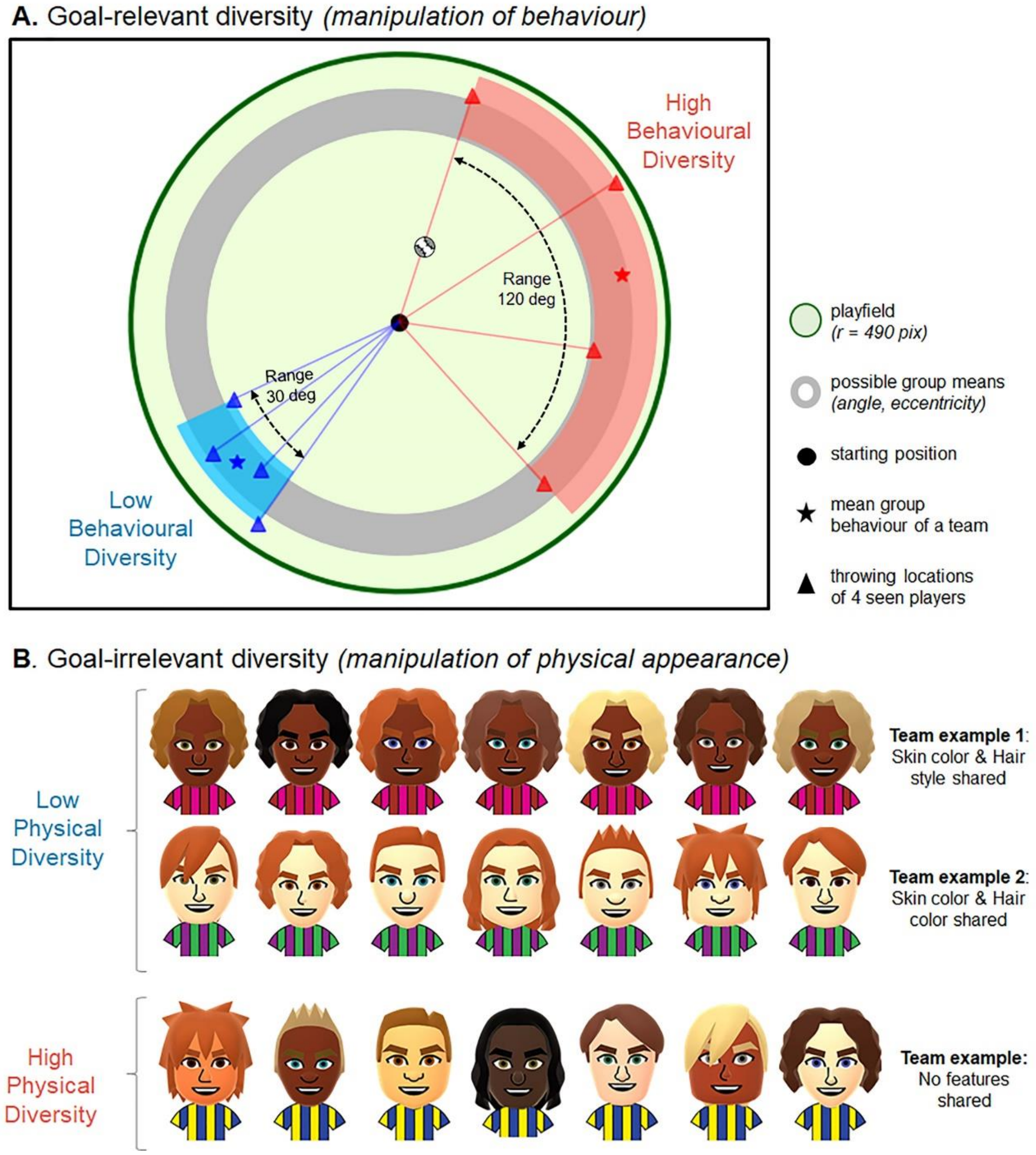

Figure 1. Manipulation of behavioural, goal-relevant diversity (A) and physical, goal-irrelevant diversity (B). A. Presented in blue is an example of a low behavioural diversity team. Presented in red is an example of a high behavioural diversity team. **B.** The two teams of low physical diversity on the top row represent the two possible counterbalancing options: half of these teams share skin colour and hairstyle, while the other half share skin colour and hair colour. The other four features remain unique across players. The team on the bottom row is an example of high physical diversity. It can be seen that each player is unique in all the six manipulated features.

Instructions. After being informed about the experiment, participants could ask questions and, if everything was clear, they gave their written informed consent, Afterwards, more detailed instructions were given. The task was phrased as a training program for the participant to become a catcher in the Catch-the-Ball game, a fictional team sport. The introductory instructions were displayed on the screen as an invitation letter to take part in this training program (i.e. the task), where the rules and phases of the game are explained. They were told that in this game one person is the catcher (that would be the participant throughout the whole game) and the other seven team members are the throwers. The throwers are the ones throwing a ball from the centre of the field, while the catcher (i.e. the participant) is the one running from the centre to catch the ball with a net. The goal as a team is that the ball gets caught with the smallest possible net, and points are given for that. They were also told that in each round, they would first meet the team members (phase 1), then observe some of them throwing (phase 2), then they would have to try to catch the ball (phase 3), and then they would be asked a question about the team (phase 4). Before starting the actual task, they underwent a step-by-step practice round with an unrelated team. Afterwards, as the participant is being trained to become a good catcher, they do this game for 28 rounds, each time with a different team.

Phase 1: Meeting the team. During the first phase of a round, the seven team members are first displayed simultaneously on the screen. Participants are told that, during this phase, the coach of the team is saying to all the throwers (but not to the catcher, i.e. the participant) where they should try to throw the ball during this round. After all the throwers secretly received these instructions, participants were told that four of the team members get randomly selected to participate in the second phase of the game. These 4 members are marked with a square so that the participant can see who will be involved in the next phase. Although this is not mentioned to the participants, in this phase our physical diversity manipulation becomes evident, as participants have the opportunity to see the appearance of all team players at once.

Phase 2: Observation. Participants are told that during the second phase of a round, they will have the chance to observe where some of the team members throw the ball. That is because the catcher (i.e. the participant) does not know where the coach told the players to throw the ball. Participants are instructed to observe where the four throwers that got randomly selected throw the ball. They are also told that later they will have to try to catch the ball, but that for now they only have to observe. First, one of the selected throwers appears on the centre of the playfield with a ball and throws the ball at a certain eccentricity and angle from the centre, according to a pre-defined manipulation of behaviour explained later. Subsequently, the next player appears and throws the ball, until the four of them have shown to the catcher how they throw. These four throwers that participants observe in phase 2 are referred to as the seen players. The three team members that have not been observed during this phase are then the unseen players. Importantly, the diversity of the throwing behaviour of the team becomes evident to the participant in this observation phase. If the team has low behavioural diversity, the throws of the four seen players will be closer together (within a range of 30°), whereas if the team has high behavioural diversity, their throws will be further away from each other (within a range of 120°). Throughout this phase, the participant not only has the opportunity to form an individualized impression of where each player throws the ball (individual knowledge) but also a group-level impression of how the team generally throws the ball (group knowledge), in terms of the mean location across throws, and their diversity. We are interested in how group diversity in throwing behaviour (goal-relevant) and physical appearance (goal-irrelevant) influence this group knowledge.

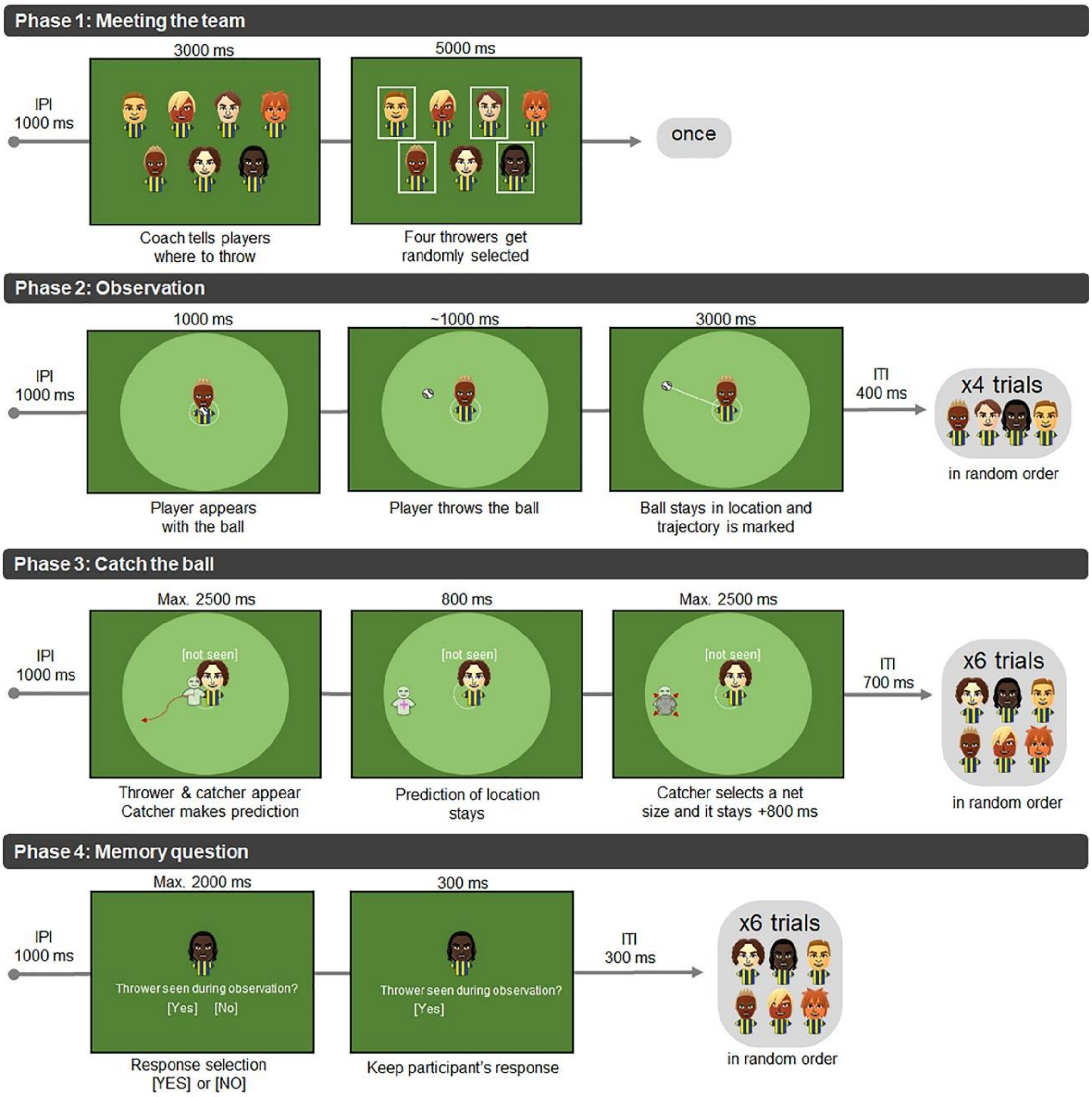

Figure 2. Schematic timeline of one round in the Catch-the-Ball game, from phase 1 to 4. The displayed stimuli such as the team players, the catcher and the ball correspond to the real stimuli used in the experiment (but they are rescaled for visualization purposes). IPI stands for inter-phase-interval, while ITI stands for inter-trial-interval.

Phase 3: Catch the ball. For this phase, participants are told that the main objective is that the ball gets caught. Participants, as the catcher of the team, will do that by running towards the predicted end location of the thrown ball and using a net to subsequently catch the ball. Participants are asked to make predictions about where six team players might throw the ball, one after another. Participants are told that of these six players, half are team members they have not observing throwing the ball before, (the unseen players). The remaining three players are members they observed in the previous phase, thus seen players. To ensure that participants know which of these team members are the previously seen ones and which ones are the previously unseen ones, a small tag saying [seen] or [unseen] was displayed on top of the thrower when a prediction had to be made.

First, the participant (i.e. the catcher) and one of the throwers of the team get ready at the centre of the playfield. As soon as the thrower becomes visible, the participant is instructed to have the catcher run as fast as possible to the location where they think the current thrower might throw the ball. To move the catcher (i.e. themselves), they can move the mouse around the field and click on the location where they predict the current player would throw at. They have a time limit of 2.5 seconds to make a prediction. The prediction, i.e. the selected location, is recorded in polar coordinates. Once they selected a position (or they ran out of time), they could not move the catcher anymore and a small net (starting net diameter = 7.5 px) appears at the centre of the catcher. This net needs to be opened in order to actually catch the ball. They are instructed to adjust the size of the net depending on how certain they are that the current player will throw the ball to that specific location. That is, they should select a smaller net size if they are very certain about their prediction (smallest net diameter = 0 px), whereas they should select a larger net size if they are less certain about their prediction (largest net diameter = 150 px). They are instructed to use the scroll wheel of the mouse and to wheel upwards (larger) or downwards (smaller) to adjust the net size. To confirm the chosen net size, participants had to click the left mouse button. The available time to select a net size was limited to 2.5 seconds. The selected net size based on the diameter (in pixels) was recorded.

Importantly, after participants placed their prediction and selected the appropriate net size to state their confidence, no feedback was given about the location where the current thrower eventually threw at. In other words, participants did not observe where the throwers actually throw at and whether or not they caught the ball. This was done because we are only interested in their predictions, and did not want participants to use the feedback to adjust their predictions accordingly. Participants, however, were told that no feedback would be provided but that the points are being counted "in the background" according to the precision of their catch; and that the total points will be revealed to them at the end of the game. However, none of the players is actually throwing a ball "in the background", so no real points are being counted.

Phase 4: Memory question. During the last phase of a round, participants are asked to report which players were the ones that took part in the second phase (the seen players) and which players did not take part in that phase (the unseen players). They are asked, one after another, about each of the six players they just made predictions about. Per thrower, they have to answer the question "Thrower seen during observation?". They had to press [yes] if they think this thrower was a seen player, and [no] if they think it was an unseen player, a time limit of 2 seconds was given for each player. If no response was given, it would automatically ask about the next thrower. The goal of this phase is to encourage participants to pay attention to the identity of the team members and to reduce the likelihood of confusing them with each other. The data from this task is mainly used to assess to what degree participants learnt who was who within each team, and to what extent they would confuse them in memory. Additionally, it was also used to assess whether the difference between high and low group diversity physical appearance was sufficient.

General. A full round with a single team typically lasts around 90 seconds. Participants could decide when to take a break while doing the task, as they had to actively press a button after each round to start with the next one. Each participant performed a total of 28 rounds of this game, with a different team each time. The task lasted a total of around 40 minutes. When the task was completed, all participants received the same feedback regarding their score on the game (which was a fixed score of 83,6/100 points).

2.4. Manipulation of diversity in behaviour and physical appearance

Group diversity is manipulated in terms of goal-relevant and goal-irrelevant diversity. Given that in this task the predictions to be made are about the throwing behaviour of the players, behavioural diversity corresponds to goal-relevant diversity (figure 1A), while physical diversity (in terms of appearance) corresponds to goal-irrelevant diversity (figure 1B).

For each team, we first define the average throwing location where the players would throw at. This is subsequently referred to as the group mean. To define the group mean in terms of its angle and eccentricity, a uniformly random angle between 0 and 359 degrees and a uniformly random eccentricity between 353 and 431 pixels are selected. Each of the four team members throws the ball to a location nearby the group mean so that the average of the four throws corresponds to this point, both in terms of angle and eccentricity. To determine where each individual player throws, a range of eccentricities and angles around the group mean is selected. A schematic representation of these elements is shown in figure 1A.

Behavioural diversity is manipulated by changing the range of angles at which individual team members would throw at. The range of eccentricity, however, is kept constant across teams and conditions. For a team of low behavioural diversity, the angular range is set to 30°, whereas for a team of high behavioural diversity, the range is set to 120° (see figure 1A for examples). Therefore, individual players throw the ball closer to the group mean when they belong to a low diverse team compared to a highly diverse team. The angles at which the four team members throw are determined by selecting four roughly equally spaced points within a range of 30° or 120° around the angular mean of the group. Similarly, the eccentricities at which the four team members throw are determined by selecting four roughly equally spaced points within a range of 116 px around the eccentricity mean of the group. These four angles and four eccentricities are then randomly combined, providing four throwing locations in polar coordinates, randomly assigned to the four selected throwers. This randomization allows us to avoid showing the same distribution of throws in space across teams. Importantly, this procedure ensures that the average throwing location corresponds to the group mean without none of the players actually throwing to the exact group mean. In other words, the average throwing location of a team could not be observed but had to be inferred.

Physical diversity is manipulated by changing how similar the appearance across group members is in terms of skin colour and hair (see figure 1B for examples). For low physical diversity teams, members looked very similar to each other. Particularly, they shared two features (skin colour and hair colour/style) and were unique on the other 4 features (hair colour/style, nose shape, eye colour and face shape) (so that they are still distinguishable). For instance, all team players could have the lightest skin colour and curly long hair, while having each of them a unique nose, hair colour, face shape and eye colour. Notice that, for low physical diversity teams, hair colour and hairstyle were counterbalanced across teams and conditions to be sometimes a shared feature and sometimes a unique feature. In contrast, for high physical diversity teams, members looked very diverse as each member was unique for all physical features. That is, all member had a different phenotype for each of six different features (unique skin colour, hair colour, hairstyle, nose, eye colour and face shape), and did not share any of them. The manipulations to create low and high physical diversity were meant to make the difference between them clear enough while still being able to distinguish the team players within a low diversity team. Note that we choose to manipulate diversity specifically in skin colour and hair (colour or style, depending on the counterbalancing) so that it is more salient That is, differences in skin colour are more noticeable than differences in the nose shape. Moreover, diversity in skin colour or hair is more plausible to encounter in real-life groups (e.g. a multinational group can be diverse in skin colour, a hippie group can be similar in hairstyle).

Behavioural and Physical diversity are combined in a 2 x 2 design, leading to the following 4 possible team configurations: (i) team members display more similar behaviour and look similar, (ii) team members display similar behaviour and look diverse, (iii) team members display diverse behaviour and look similar and (iv) team members display diverse behaviour and look diverse. There are 7 teams per condition, thus 28 teams in total.

2.5. Dependent variables and data pre-processing

Throughout the task, several dependent measures were recorded. Regarding the predictions for the three seen and three unseen players during the third phase of a round, we recorded the polar coordinates of each of these predicted throwing locations and the uncertainty of these predictions in terms of the selected net size. Moreover, we recorded the percentage of correct answers in the memory question of the last phase of each round.

Based on these recorded measures, several variables have been computed to assess whether predictions are more variable, inaccurate and uncertain as a function of group diversity. Since we are interested in how the predictions about unknown members' behaviour are influenced by the different diversity manipulations, we only analysed the predictions made for the three unseen throwers. This is because only these players are the ones for which participants only have group knowledge about, but no individual knowledge. Another aspect we take into account is that behavioural diversity was manipulated at the angular level but not in terms of eccentricity. Therefore, the computed variables related to the predictions are calculated based on the angular value (in degrees) of their polar coordinates, ignoring the eccentricity value. This data was then used to compute the variables stated below. Data pre-processing was performed using custom code in Python 3.4.

Range of predictions. To investigate the effect of group diversity on the variability of predictions about the behaviour of unseen team members, we calculated the angular range of the 3 predictions. This will be referred to as the range of predictions. To compute this variable, we took from the three predictions the highest and the lowest angle and compute the absolute distance between them. This was done in such a way that the wrap-up at 360° was taken into account. This angular range was then averaged across the teams within each of the four conditions and is reported in degrees. A larger range of predictions indicates more variable predictions.

Single-level mean error. The second aspect to be addressed is the extent to which predictions diverge from the average group behaviour, thus the inaccuracy of predictions in relation to this reference. We assess this aspect by means of the error or the distance between single predictions and the group mean of a team, which we refer to as the single-level mean error. To compute this variable, first, the absolute angular distance between each of the three predictions and the group mean is calculated. Taking the average of these three absolute distances subsequently leads to the single-level mean error. This variable was then averaged across the teams within each of the four conditions and is reported in degrees. The single-level mean error reflects how far on average a single predictions are placed relative to the average behaviour of the group. In other words, single-level mean error can be interpreted as the extent to which the central tendency of the group is used when making a single prediction The higher its value, the more inaccurate and divergent from the mean a single prediction is.

Net size. To assess the uncertainty associated with each prediction, the selected net size (diameter in pixels) per each prediction was recorded. The larger the size of the net, the more uncertain the participant was about a particular prediction. The average net size used for the three predictions of unseen players of one round is calculated. This variable was then averaged across the teams within each of the four conditions and is reported in pixels.

Finally, the accuracy during the memory phase of each round, where participants were asked about six players, was recorded in the form of a percentage of correct answers. This memory accuracy was averaged across teams of the same condition and then compared across conditions. If this variable is above the chance level of 50%, it can be interpreted as participants being able to distinguish seen and unseen members.

2.6. Data analysis

This study follows a 2 (behavioural diversity: low vs. high) by 2 (physical diversity: low vs. high) full within-subjects factorial design. Analysis was done by using R software version 3.5.3 and the *ez* package (Lawrence, 2016), All dependent variables (range of predictions, single-level mean error, net size and memory accuracy) were subjected to a 2-by-2 repeated measures ANOVA (rmANOVA) with Behavioural Diversity (Low, High) as one within-subjects factor and Physical Diversity (Low, High) as the other within-subjects factor. Whenever an interaction effect was found, post-hoc paired-sampled t-tests were used to further analyse simple effects across conditions. Specifically for memory accuracy, one-sampled t-tests were used to assess whether accuracy was above chance level (50%). All given *p*-values are two-tailed. The mean and standard deviation reported per any given condition refer to the descriptive statistics of the values under that condition, without taking into account the within-subjects design. However, the standard error (SE) displayed on the graphs are corrected for the within-subjects nature of our data.

The same analysis excluding the participants that overall did not generalize the main team behaviour to unseen team members was carried out. To evaluate this, we looked at the distance between the average of the participant's predictions and the mean behaviour of the group (referred to as generalization error). To compute the generalization error, first, the average angle from the three predictions is calculated. This was done by taking into account the wrap-up at 360° and making sure that the average angle is somewhere within the shortest range of the predictions. Then the absolute angular distance between this average prediction and the group mean is computed, yields to the generalization error of a team. This variable is averaged across the teams of each condition and then averaged across conditions. This leads to a single value per participant, which is their average generalization error (reported in degrees). The average generalization error indicates the extent to which the average prediction for unseen members was close to the mean behaviour of the group, regardless of condition. The mean and standard deviation of this variable were computed across participants (M = 26.05°, SD = 17.73°). A participant was considered an outlier if their average generalization error was above 2.5 standard deviations from the mean, which is above 70.39°. In that case, participant's predictions were centred around a point too far from the average group behaviour they observed. This could be interpreted as if the participant discarded unseen members as not being part of the team and not following the instructions from the coach (who tells all players to try to throw to a certain location). For this study, an overall generalization effect was necessary and expected across conditions in order to evaluate differences across conditions due to diversity manipulations.

Following these criteria, three participants were detected as outliers (4.6% of the total), hence the additional analysis without outliers included 62 participants. This additional analysis consisted of exactly the same dependent variables, factors and tests as the main analysis. A complete report of this analysis without outliers is included in the Supplementary Materials. However, when a result of the analysis without outliers is qualitatively different from the result of the analysis including all participants, mention is made in the main text. A qualitative difference refers to cases in which the significance (or trend) and or the direction of an effect differs between both analyses.

3. Results

First, the rmANOVA for the range of predictions (see Figure 3A) revealed a main effect of Behavioural Diversity ($F(1, 64) = 120.74$, $p < .0001$, $\eta_p^2 = .65$). Particularly, the range of predictions was larger when the team's behavioural diversity was high (M = 75.70°, SD = 34.39°) compared to when it was low (M =

43.21°, *SD* = 38.36°). Although no significant main effect of Physical Diversity was found ($F(1, 64) = 2.48$, $p = .12$), a trend in the hypothesized direction was descriptively noticed, namely, a larger range of predictions for high physical diversity (*M* = 60.77°, *SD* = 35.58°) compared to low (*M* = 58.14 °, *SD* = 34.57°). In fact, the analysis without outliers (see figure S1) did reveal a main effect of Physical Diversity ($F(1, 61) = 5.04$, $p = .03$, $\eta_p^2 = .08$). That is, the range of predictions was larger when the team's physical diversity was high (*M* = 58.64°, *SD* = 34.77°) compared to when it was low (*M* = 54.98°, *SD* = 31.99°). Finally, no interaction effect was found ($F(1, 64) = 0.98$, $p = .33$).

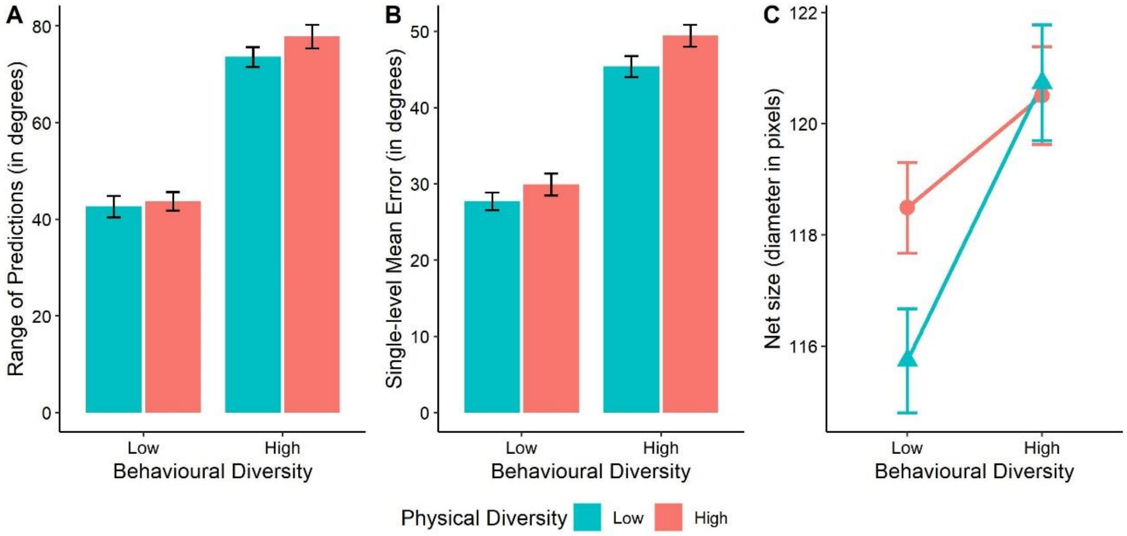

Figure 3. Effects of Behavioural Diversity and Physical Diversity on (A) range of predictions, (B) Single-level mean error and (C) Net size. Error bars represent the SE corrected by within-subjects design. The data from all participants (N = 65) is depicted in these plots.

Second, the rmANOVA for the single-level mean error (see figure 3B) revealed a main effect of Behavioural Diversity ($F(1, 64) = 123.16$, $p < .0001$, $\eta_p^2 = .66$). Particularly, single predictions were placed on average further away from the group mean when the behavioural diversity of the team was high (*M* = 47.41°, *SD* = 32.18°) compared to when it was low (*M* = 28.81°, *SD* = 31.33°). Moreover, a main effect of physical diversity was found ($F(1, 64) = 8.09$, $p = .006$, $\eta_p^2 = .11$). Particularly, single predictions were placed on average further away from the group when physical diversity of the team was high (*M* = 39.66°, *SD* = 32.62°) compared to when it was low (*M* = 36.56°, *SD* = 30.01°). No interaction effect was found ($F(1, 64) = 0.56$, $p = .46$).

Third, the rmANOVA with net size as dependent variable (see figure 3C) revealed a main effect of Behavioural Diversity ($F(1, 64) = 10.71$, $p = .002$, $\eta_p^2 = .14$).

Particularly, a larger net size was selected when the team's behavioural diversity was high (M = 120.6 px, SD = 31.34 px) compared to when it was low (M = 117.1 px, SD = 31.88 px). No main effect of physical diversity was found ($F(1, 64) = 1.93$, p = .17). However, a trend towards an interaction effect was found ($F(1, 64) = 3.85$, p = .05, $\eta_p^2 = .06$). A post-hoc paired t-test revealed a simple effect of low behavioural diversity ($t(64) = 2.56$, p = .01). Particularly, the selected net size was smaller for low physical diversity (M = 115.7 px, SD = 31.76 px) than for high physical diversity (M = 118.5 px, SD = 32.58 px), when behavioural diversity was low. In contrast, no simple effect for high behavioural diversity was revealed by paired t-test ($t(64) = -0.17$, p = .86). That is, when the team's behavioural diversity was high, physical diversity being low (M = 120.7 px, SD = 31.64 px) or high (M = 120.5 px, SD = 31.89 px)did not make any difference on the selected net size.

Finally, we looked at memory accuracy during phase 4. Overall mean accuracy during the memory task was 78.1 % (SD = 10.39 %), which is significantly above chance level (50%), $t(64) = 21.77$, $p < .0001$. The rmANOVA for memory accuracy revealed a main effect of Physical Diversity ($F(1, 64) = 37.75$, $p < .0001$, η_p^2 = .37). Particularly, participants could categorize more accurately which players were seen or unseen when the team's physical diversity was high (M = 80.82 %, SD = 11.45 %) compared to when it was low (M = 75.31%, SD = 10.54%). Neither a main effect of Behavioural Diversity ($F(1, 64) = 0.25$, p = .62) nor an interaction ($F(1, 64)$ = 0.01, p = .92) were found.

4. DISCUSSION

The present study aimed to investigate how predictions about the behaviour of unknown group members are affected by goal-relevant group diversity (i.e., behavioural diversity) and goal-irrelevant group diversity (i.e., physical diversity). In our task, both behavioural and physical diversity were manipulated and participants had to make predictions about the behaviour of unknown group members. As expected, we found that when unknown group members belong to a group of high behavioural diversity, predictions about their potential behaviour are more variable, more inaccurate relative to the group mean behaviour, and more uncertain, as compared to a group of low behavioural diversity. Importantly, our results also show that these predictions are affected by physical diversity, a dimension that, in this context, is unrelated to the behaviour of the group members. In particular, physical diversity led to slightly more variable and more inaccurate predictions relative to the group behaviour, as well as more uncertainty around these predictions. Taken together, our findings suggest that when we encounter unknown members of a group, our predictions regarding their behaviour are influenced not only by how diverse this behaviour usually is within the group they belong to, but also by how diverse group members are on a different, unrelated dimension, which in the present study is physical appearance.

Concerning goal-relevant diversity, we found that higher variability in the relevant group property leads to (i) predictions that are more variable among unknown group members and (ii) individual predictions that are further away from the average behaviour of the group. This goes in line with previous studies on goal-relevant group diversity (Dasgupta et al., 1999; Lambert, 1995; Park & Hastie, 1987). These results indicate that we expect unknow members to behave more different to each other and more divergent from the mean group behaviour when the group they belong to happens to behave diversely in that same dimension (in this case, throwing a ball during sports). Notably, in order to make informed predictions, participants had to form an impression of the behaviour of the group. As discussed in previous research, group knowledge regarding a specific property is not only characterised by its central tendency but also by how this property is distributed within the group, which includes the variability of this property across group members (Fried & Holyoak, 1984; Tenenbaum & Griffiths, 2001; Sakamoto et al 2008; Osherson 1990). Earlier research has already shown that group knowledge is applied to group members (Allport, 1954; Fiske & Neuberg, 1990; Macrae & Bodenhausen, 2000; Macrae & Bodenhausen, 2001). Given that the variability of a group property is an important part of group knowledge, it is reasonable to expect that this diversity will be applied to group members as well (cf. Park & Hastie, 1987; Lambert, 1995). Our results show that this is the case: group diversity, as part of group knowledge, is applied to newly encountered group members itself. In particular, higher group variability in a certain dimension also leads to more variable predictions on that same dimension. Moreover, these predictions are further away from the central tendency of the group. Finally, our results show that these predictions are associated with more uncertainty when behavioural diversity is high (Lambert et al., 1998; see also Kearney, 2011). This is an interesting effect in light of Ryan et al. 's findings (1996). Their study showed that perceived diversity affects the certainty of judgments, but not the judgment itself. In our study, however, the effect of group diversity was not limited to the uncertainty with which we make predictions; but it directly affected the variability of the predictions themselves.

Regarding goal-irrelevant diversity, we found that the level of group diversity in terms of physical appearance does influence predictions about unknown group members' behaviour. Particularly, when physical diversity within a group is high, predictions about unknown group members' behaviour become more variable and more divergent from the average group behaviour. This outcome resembles previous findings (Abelson et al., 1998; Dasgupta et al., 1999). Although the effect on variability was only significant in the analysis without outliers, the same trend was observed in both. Regarding uncertainty, the results suggest that higher physical diversity increases uncertainty about predictions, but only when the behavioural diversity was low. We argue that this interaction could be due to a task-related ceiling effect. Participants had to scroll up the mouse to open the net to indicate uncertainty, thus a larger net required more scrolling. Since there was limited time to select a net size, maybe participants did not have time to scroll larger than this. Regardless of the specific reasons this effect, our findings show that goal-irrelevant diversity seems to influence prediction's uncertainty as well. In general, these effects go in the same direction as the effects of goal-relevant diversity.

As mentioned earlier, diversity in physical appearance, particularly in this task, has no informative value to make predictions about member's throwing behaviour. Moreover, it is difficult to imagine how higher physical diversity could be semantically associated with the goal of the task (e.g., multi-ethnic teams tend to throw the ball to different locations). Interestingly, Goldstone & Son (2012) argued that if group knowledge about the property you want to make predictions about is available, this relevant information is taken into account over uninformative one. Our results show that this is not necessarily the case. In contrast to this, we found that goal-irrelevant diversity did influence predictions over and above goal-relevant diversity. Importantly, this happened even though the goal-irrelevant knowledge coexisted with group knowledge on a goal-relevant dimension. How can we then explain that variability on a goal-irrelevant dimension nevertheless affects predictions? We suggest three possible mechanisms. First, in the field of group impression formation, it has been argued that physical similarity (or any other similarity within the group) leads to perceiving a social group more as a whole (Dasgupta et al., 1999; Crawford et al., 2002; Nelson, 2009). It is then this perception of groupness, regardless of its origins, that leads to judging members as more similar to each other and leads to more confident judgments. A second possible explanation could be in the lines of typicality-based inference. It has been shown that it is harder to assess typicality when the group is diverse (Lambert, 1995). Therefore, in the high physical diversity groups, it is probably harder to assess what a prototypical member looks like. As a consequence, any member of this group could be seen as less prototypical, or their typicality might be unclear. A recent study has shown that if a member is prototypical relative to the group, inferences are more based on group knowledge (Ma, Cornell & Wittenbrink, 2018). A third explanation could be related to low-level processes. Group members of a low physical diversity team share some visual features (namely, skin colour and hair), and thus the behaviour of group members could have been associated to these visual features they have in common, rather than to them as individuals or to the social notion of a group. In the case of high physical diversity groups, members shared no feature apart from the team t-shirt. Thus, in this case, common visual features to which behaviour could be associated are missing, in comparison to low physical diversity groups. Taking these three theories together, perhaps a lower sense of groupness (i), less prototypicality in appearance (ii) and/or the lack of shared visual features (iii) led to making more variable and uncertain predictions about unknown members' behaviour.

In relation to this, it could be argued that participants might have confused unseen and seen players and this is why the predictions for unseen players were more similar. However, there are a number of reasons why this is rather unlikely: (i) memory accuracy was above chance level, also for low physical diversity teams, (ii) there was a visual reminder about players being seen or unseen and (iii) they knew every prediction was about a different player. Therefore, it seems unlikely that the reason behind the physical similarity effect lies on a confusion about the group members; but rather on aspects of the suggested explanations.

Although our results are generally in line with previous studies in the field, this study overcomes previous methodological limitations and extends our knowledge on the effects of group diversity. First, this is the only study, to our knowledge, that has manipulated group diversity on a relevant and an irrelevant dimension simultaneously. This allowed us to study not only the effects of each of them but also their effect on coexistence, as discussed above. Second, in this study participants made fast behavioural predictions during an interactive computer task. Hence, these predictions were more implicit and meaningful for the participant compared to the explicit measures used in previous research (Abelson et al., 1998; Dasgupta et al., 1999; Kim & Lee, 2017; Lambert, 1995; Park & Hastie, 1987; Patalano & Ross, 2007; Ryan et al., 1996), which enhances the ecological validity of our study. Similarly, in our task participants were empirically exposed to the individual behaviours of group members and their appearance. By aggregating these empirical observations they could infer group knowledge, which resembles how we estimate group characteristics in real life. This is a methodological improvement in comparison to previous studies that used explicit descriptions to induce group knowledge and perceptions of diversity (e.g. Crawford et al., 2002; Patalano & Ross, 2017) or made use of pre-existing groups (e.g. Ryan et al., 1996; Kim & Lee, 2017). Finally, we assessed multiple aspects of predictions to look at the effects of group diversity (variability, inaccuracy and uncertainty). The effects of goal-relevant and goal-irrelevant diversity are not necessarily the same on all these aspects. Thus, having extensive coverage of the possible effects allowed us to get a better picture of how these two different types of diversity influence predictions of unknown groups members.

This study addresses the mentioned challenges present in previous research and brings in new insights. Yet, there are a few limitations that should be pointed out, and potentially addressed in future research. First, the throwing behaviour assigned to group members was quantitatively manipulated, could be observed in space and did not have any positive or negative valence. It can be argued that such neutral group knowledge does not really resemble the type of group knowledge we usually have about social groups. Namely, group knowledge usually concerns rather abstract, qualitative aspects such as psychological traits, and usually, it has a certain valence (i.e. a group being smart is a better property than lazy). Yet, abstract properties with valence can also be more or less variable within a group and could have an approximate central tendency. Excluding abstract properties and valence from our study allowed us to actually have more experimental control over the effects and thus isolate the most neutral and basic underlying process. The fact that we found the discussed effects despite the neutrality of our manipulation, makes it seem likely that our findings would still be found with more realistic group properties. Yet further work should be done in this direction. Second, we asked participants to make predictions about multiple new group members individually. Perhaps if they had to make a single prediction (either about one single new player or new players as a whole), variability would not have affected the prediction itself. As Ryan et al. (1996) showed, variability can lead to more uncertainty but not necessarily to a change in the judgment itself. However, since we asked participants to make predictions about multiple new members, group diversity was probably used to distribute the expected behaviour variability across members. Future studies could investigate how the effect of group diversity on predictions is affected by the number of predictions to be made. Third, we manipulated goal-irrelevant diversity in terms of physical appearance, particularly in skin colour and hair. Notably, this group property has certain connotations in real-life groups and it is immediately available to participants when encountering a new group member. To control for these two possibly confounding aspects, future studies could manipulate goal-irrelevant diversity for instance on physical features that are more neutral (e.g. nose shape or eye colour), on a property that is not directly available (e.g. throwing trajectories of players) or even using non-social stimuli. Finally, we used time-limited meaningful predictions to assess the effects of group diversity in our task. Although this is a more implicit and meaningful assessment in comparison to previous methods used in the field, additional measures could be used in future studies. For instance, to capture group diversity effects more implicitly, electrophysiological responses to expectation violation while observing group members' behaviour could be used. As for the assessment of group diversity effects in a more embedded and interactive setting, the game designed for this task could be actually done with real or virtual agents.

Our findings have important implications for real-life situations. For instance, even when we have group knowledge about a property ("I know teenagers are usually late…"), diversity in another dimensions that has no predictive value ("…but teenagers have very different hairstyles,") could influence our predictions about unknown group members ("so perhaps I'm not sure anymore whether Maria will be late"). In relation to this, Brauer & Er-rafiy (2011) found that increasing perceived diversity within a group seems to reduce the strength and the certainty with which a stereotype (i.e. group knowledge) is applied to group members, which is also supported by our findings. A possible application of this idea could be to actively increase diversity within groups of co-workers, classrooms or sports teams, both in relevant and irrelevant domains. For instance, increasing diversity within a classroom in terms of academic performance (relevant) but also in terms of gender, ethnicity or aesthetic styles (theoretically irrelevant), could help teachers make fewer assumptions about pupils' academic performances and be more open to diversity within the group.

To conclude, we have provided evidence that both goal-relevant and goal-irrelevant diversity influence our predictions about unknown group members. Particularly, higher group diversity leads to more variable, inaccurate and uncertain predictions. Our findings suggest that goal-irrelevant diversity (here, in physical appearance) influences our predictions over and above the effects of group-relevant diversity (here, in behaviour). After discussing possible underlying mechanisms, strengths and limitations of the current study, we suggest potential future research directions and societally relevant applications.

5. References

Abelson, R. P., Dasgupta, N., Park, J., & Banaji, M. R. (1998). Perceptions of the collective other. *Personality and Social Psychology Review*, *2*(4), 243-250.

Allport, G. W. (1954). *The nature of prejudice*. Reading, MA: Addison-Wesley.

Axelson, R. D., Solow, C. M., Ferguson, K. J., & Cohen, M. B. (2010). Assessing implicit gender bias in medical student performance evaluations. *Evaluation & the health professions, 33*(3), 365-385.

Bach, P., & Schenke, K. C. (2017). Predictive social perception: Towards a unifying framework from action observation to person knowledge. *Social and Personality Psychology Compass*, *11*(7), e12312.

Biernat, M., Collins, E. C., Katzarska-Miller, I., & Thompson, E. R. (2009). Race-based shifting standards and racial discrimination. *Personality and Social Psychology Bulletin, 35*(1), 16-28.

Brauer, M., & Er-Rafiy, A. (2011). Increasing perceived variability reduces prejudice and discrimination. *Journal of Experimental Social Psychology*, *47*(5), 871-881.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, *36*(3), 181-204.

Crawford, M. T., Sherman, S. J., & Hamilton, D. L. (2002). Perceived entitativity, stereotype formation, and the interchangeability of group members. *Journal of personality and social psychology*, *83*(5), 1076.

Dasgupta, N., Banaji, M. R., & Abelson, R. P. (1999). Group entitativity and group perception: Associations between physical features and psychological judgment. *Journal of personality and social psychology*, *77*(5), 991.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In *Advances in experimental social psychology* (Vol. 23, pp. 1-74). Academic Press.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(2), 234.

Frith, C. D., & Frith, U. (2006). How we predict what other people are going to do. *Brain research*, *1079*(1), 36-46.

Goldstone, R. L., & Son, J. Y. (2012). Similarity. In The Oxford Handbook of Thinking and Reasoning (pp. 1?38). https://doi.org/10.1093/oxfordhb/9780199734689.013.0010

Green, A. R., Carney, D. R., Pallin, D. J., Ngo, L. H., Raymond, K. L., Iezzoni, L. I., & Banaji, M. R. (2007). Implicit bias among physicians and its prediction of thrombolysis decisions for black and white patients. *Journal of general internal medicine, 22*(9), 1231-1238.

Ito, T. A., & Bartholow, B. D. (2009). The neural correlates of race. *Trends in cognitive sciences, 13*(12), 524-531.

Kearney, L. (2011). *The role of category variability in belief revision* (Doctoral dissertation, Faculty of Science, The University of New South Wales).

Kim, S., & Lee, G. H. (2017). Influence of category coherence and type of base-rate acquisition on property generalization. *Acta psychologica*, *172*, 64-70.

Lambert, A. J., Barton, L. R., Lickel, B., & Wells, J. (1998). The influence of group variability and processing goals on the ease of making judgments about social categories. *Personality and Social Psychology Bulletin*, *24*(8), 807-820.

Lambert, A. J. (1995). Stereotypes and social judgment: The consequences of group variability. *Journal of Personality and Social Psychology*, *68*(3), 388.

Lawrence, M. A. (2016). ez: Easy Analysis and Visualization of Factorial Experiments. R package version 4.4-0. https://CRAN.R-project.org/package=ez

Ma, D. S., Correll, J., & Wittenbrink, B. (2018). The effects of category and physical features on stereotyping and evaluation. *Journal of Experimental Social Psychology*, *79*, 42-50.

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual review of psychology*, *51*(1), 93-120.

Macrae, C. N., & Bodenhausen, G. V. (2001). Social cognition: Categorical person perception. *British journal of psychology*, *92*(1), 239-255.

Macrae, C. N., Milne, A. B., & Bodenhausen, G. V. (1994). Stereotypes as energy-saving devices: A peek inside the cognitive toolbox. *Journal of personality and Social Psychology*, *66*(1), 37.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences, 109*(41), 16474-16479.

Nelson, T. D. (2009). *Handbook of prejudice, stereotyping, and discrimination*. Psychology Press.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological review*, *97*(2), 185.

Park, B., & Hastie, R. (1987). Perception of variability in category development: Instance-versus abstraction-based stereotypes. *Journal of Personality and Social Psychology*, *53*(4), 621.

Patalano, A. L., & Ross, B. H. (2007). The role of category coherence in experience-based prediction. *Psychonomic Bulletin & Review*, *14*(4), 629-634.

Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Gatenby, J. C., Funayama, E. S., Gore, J. C., & Banaji, M. R. (2000). Amygdala activation predicts performance on indirect measures of racial bias. *Journal of Cognitive Neuroscience, 12*(5), 729-38.

Quadflieg, S., Flannigan, N., Waiter, G. D., Rossion, B., Wig, G. S., Turk, D. J., & Macrae, C. N. (2011). Stereotype-based modulation of person perception. *Neuroimage, 57*(2), 549-557.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Rachlinski, J. J., Johnson, S. L., Wistrich, A. J., & Guthrie, C. (2008). Does unconscious racial bias affect trial judges. *Notre Dame L. Rev.*, *84*, 1195.

Ryan, C. S., Bogart, L. M., & Vender, J. P. (2000). Effects of perceived group variability on the gathering of information about individual group members. *Journal of Experimental Social Psychology*, *36*(1), 90-101.

Ryan, C. S., Judd, C. M., & Park, B. (1996). Effects of racial stereotypes on judgments of individuals: The moderating role of perceived group variability. *Journal of Experimental Social Psychology*, *32*(1), 71-103.

Sakamoto, Y., Jones, M., & Love, B. C. (2008). Putting the psychology back into psychological models: Mechanistic versus rational approaches. *Memory & Cognition*, *36*(6), 1057-1065.

Sebanz, N., & Knoblich, G. (2009). Prediction in joint action: What, when, and where. *Topics in Cognitive Science*, *1*(2), 353-367.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and brain sciences*, *24*(4), 629-640.

6. Supplementary materials

Results of the analysis without outliers

First, the rmANOVA for the range of predictions (see Figure S1A) revealed a main effect of Behavioural Diversity (F(1, 61) = 203.02, p < .0001, η_p^2 = .77). Particularly, the range of predictions was larger when the team's behavioural diversity was high (M = 74.53°, SD = 34.73°) compared to when it was low (M = 39.08°, SD = 33.69°). A significant main effect of Physical Diversity was found (F(1, 61) = 5.04, p = .03, η_p^2 = .08). That is, the range of predictions was larger when the team's physical diversity was high (M = 58.64°, SD = 34.77°) compared to when it was low (M = 54.98°, SD = 31.99°). Finally, no interaction effect was found (F(1, 61) = 1.02, p = .32).

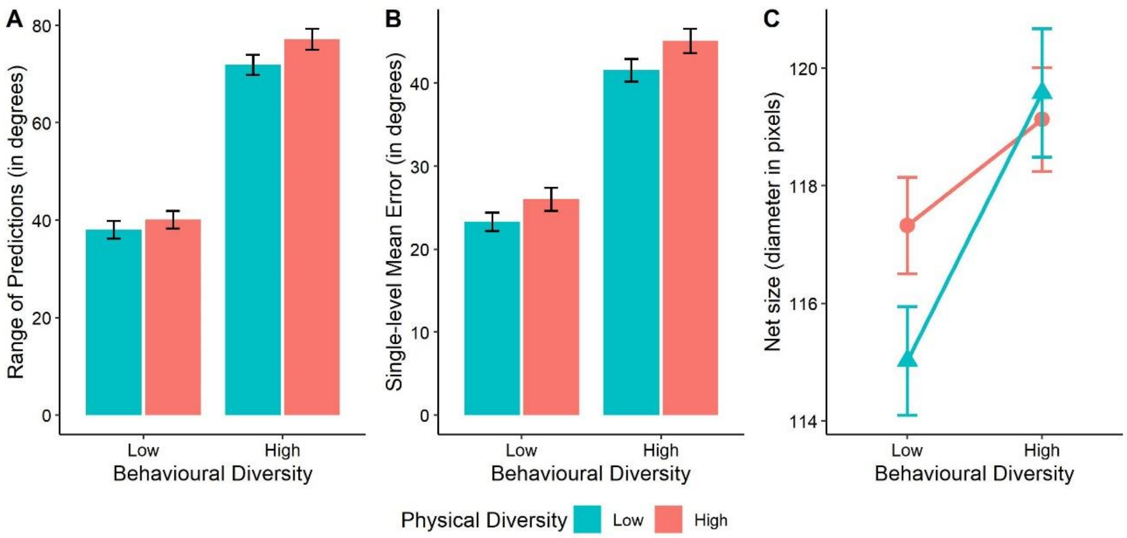

Figure S1. Effects of Behavioural Diversity and Physical Diversity on (A) range of predictions, (B) Single-level mean error and (C) Net size. Error bars represented the SE corrected by within-subjects design. Only the data from participants that were not outliers (N = 62) is depicted in these plots.

Second, the rmANOVA for the single-level mean error (see figure S1B) revealed a main effect of Behavioural Diversity (F(1, 61) = 117.47, p < .0001, η_p^2 = .66). Particularly, single predictions were placed on average further away from the group mean when the behavioural diversity of the team was high (M = 43.36°, SD = 26.87°) compared to when it was low (M = 24.68°, SD = 25.49°). Moreover, a main effect of physical diversity was found (F(1, 61) = 8.09, p = .006, η_p^2 = .12). Particularly, single predictions were placed on average further away from the group when physical diversity of the team was high (M = 35.58°, SD = 27.26°) compared to when it was low (M = 32.46°, SD = 30.01°). No interaction effect was found (F(1, 61) = 0.12, p = .73).

Third, the rmANOVA with net size as dependent variable (see figure S1C) revealed a main effect of Behavioural Diversity ($F(1, 61) = 8.50$, $p = .005$, $\eta_p^2 = .12$). Particularly, a larger net size was selected when the team's behavioural diversity was high ($M = 119.4$ px, $SD = 31.55$ px) compared to when it was low ($M = 116.2$ px, $SD = 32.31$ px). No main effect of physical diversity was found ($F(1, 61) = 1.04$, $p = .31$). However, a trend towards an interaction effect was found ($F(1, 61) = 3.10$, $p = .08$, $\eta_p^2 = .05$). A post-hoc paired t-test revealed a simple effect of low behavioural diversity ($t(61) = 2.18$, $p = .03$). Particularly, when behavioural diversity was low, the selected net size was smaller for low physical diversity ($M = 115.0$ px, $SD = 32.32$ px) than for high physical diversity ($M = 117.3$ px, $SD = 32.83$ px). In contrast, no simple effect for high behavioural diversity was revealed by paired t-test ($t(61) = -0.34$, $p = .74$). That is, when the team's behavioural diversity was high, physical diversity being low ($M = 119.6$ px, $SD = 31.94$ px) or high ($M = 119.1$ px, $SD = 32.01$ px) did not make any difference on the selected net size.

Finally, we looked at memory accuracy during phase 4. Overall mean accuracy during the memory task was 77.6 % ($SD = 10.65$), which is significantly above chance level (50%), $t(61) = 20.93$, $p < .0001$. The rmANOVA for memory accuracy revealed a main effect of Physical Diversity ($F(1, 61) = 34.61$, $p < .0001$, $\eta_p^2 = .36$). Particularly, participants could categorize more accurately which players were seen or unseen when the team's physical diversity was high ($M = 80.26$ %, $SD = 11.42$ %) compared to when it was low ($M = 74.94\%$, $SD = 10.58\%$). Neither a main effect of Behavioural Diversity ($F(1, 61) = 0.41$, $p = .53$) nor an interaction ($F(1, 61) = 0.01$, $p = .92$) were found.